Universiteit Leiden

Extending the Training Free VPint2 Framework for Multi-Temporal Cloud Removal in Satellite Image Time Series

Christiaan Moussa (s3632695)

Supervisors:
Laurens Arp & Dr. Mitra Baratchi

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 17, 2026

Abstract

Cloud cover is and will likely remain a pervasive and continuous problem for optical satellite imagery. While deep learning methods have demonstrated the ability to handle cloud removal well, they often rely heavily on very large labeled datasets and large computational resources. The training-free VPint2 algorithm has shown competitive results and performances when it comes to single-image cloud removal. It accomplishes this through the propagation of spectral information from cloud-free spatial neighborhoods using an ideally cloud-free reference image. However, the main limitation of the standard VPint2 approach is its inherently mono-temporal nature. It assumes the availability of a fully cloud-free reference image which is not always satisfied in real-world data. In this work, we extend VPint2 to the satellite image time-series setting by producing a synthetic reference image. This synthetic reference image is constructed through temporally-weighted partially cloud-free observations throughout the time-series, rather than relying on a single fully cloud-free timestep. Experiments were conducted on most of the continents present in the Sentinel-2 SEN12MS-CR-TS dataset and cloud detection was performed using the SEnSeIv2 algorithm. In this work, we used various metrics to evaluate the model's performance against the established baselines DSen2-CR, UnCRtainTS, STGAN, CR-TS Net and U-TAE. Similarly, we will also compare the performance of the extended model to the original VPint2 model. Overall, the extended VPint2 model was shown to perform competitively in comparison to the other methods and handled diverse landscapes well.

Contents

Acknowledgments

I would like to thank my supervisors Laurens Arp and Dr. Mitra Baratchi for their guidance and assistance throughout the course of the semester. From the implementation to receiving helpful feedback, the entire process has been very smooth and efficient due to both of my supervisors proactively assisting my progress.

1 Introduction

Satellite imagery is essential when researching and investigating Earth phenomena and relevant observations. Such applications include providing active fire data [1] and tracking land use for agricultural production [2]. However, using such imagery is not without hurdles and obstacles. Cloud obstruction is a major problem and limitation for satellite imagery. Seeing as clouds cover on average 67% of Earth's surface at any time frame, the usability of Earth's satellite imagery is reduced [3]. To combat this, some downstream users rely on pre-trained deep learning models for cloud removal. These models often require large datasets and specific input conditions [4].

Training-free methods avoid model training that is data intensive. One such method is VPint2 (Value Propagation Interpolation 2) [4]. VPint2 builds upon spatial interpolation principles in order to reconstruct cloudy pixels using the information derived from a cloud-free reference image alongside non-cloudy pixels. This method compares the values of direct spatial neighbors and subsequently propagates information from valid regions into the areas which are covered by clouds. This method achieves this without relying on or requiring model training. A limitation of the standard VPint2 method is the mono-temporal nature of the reconstruction. The cloud-free reference image is a crucial part of the VPint2 reconstruction. The reference image itself is based on an observation of a single time-step within the entire time-series. For these reasons, VPint2 is mono-temporal in some ways as opposed to truly multi-temporal. The concept of a multi-temporal setting used in this thesis primarily points towards the usage of multiple observations from multiple time-steps across the entire time-series. Another limitation of the VPint2 approach is its reliance on a fully cloud-free observation in the time-series. VPint2 requires both one cloudy target image and one fully cloud-free reference image to be available. In practice however, it can be challenging to find such cloud-free reference images.

On the other hand, finding time-series of partially cloudy observations is far more common. Datasets such as SEN12MS-CR-TS [5] exemplify this by providing multi-modal and multi-temporal satellite image sequences with varying cloud coverage. While deep learning methods were shown to perform very well on such datasets, these models often require extensive computational power during training [4]. Training-free methods also avoid dataset representationality issues or conflicting patterns. From this, we can see that training-free cloud removal methods that operate on multi-temporal satellite image time-series remain relatively underexplored. This work aims to address and bridge this gap in the literature by extending the VPint2 method to a multi-temporal setting.

The research question addressed in this thesis is: "How can the training-free VPint2 method be extended to perform multi-temporal cloud removal in satellite image time-series?" This research question extends the existing VPint2 literature by focusing on broadening the applicability. This is achieved by enabling VPint2 to function with realistic satellite time-series without relying on a fully cloud-free image.

Our contributions presented in this thesis are as follows:

- We propose an extension of the original VPint2 method to the satellite image time-series setting by producing a synthetic reference image. Through this synthetic reference image, our extension does not rely on a single fully cloud-free reference image in the time-series.

- We tested our extension of the original VPint2 method on nearly the entire Sentinel-2 portion of the SEN12MS-CR-TS dataset. This dataset covered the regions of Europe, Africa, Americas,

West and East Asia. This allowed for our extension to be tested on diverse landscapes, environments and cloud coverages discrepancies.

- We tested our extension of the VPint2 method against other existing multi-temporal methods alongside the original VPint2 method. We used various metrics to perform consistent comparisons and found that our extended model performed competitively and even outperformed a few of the existing methods in some metrics.

In this bachelor's thesis, we examine ways to extend the VPint2 method to multi-temporal settings. Section 1 gives a brief introduction of the problem. Section 2 provides the theoretical understanding required to perform the extension in this thesis. Section 3 will provide other models and theories discussed in relation to the problem. Section 4 lays out the general approach and steps taken on a quantitative and conceptual level. Through this, we can have a reproducible approach. Section 5 shows the experimental setup that will be employed in order to achieve the desired results. Section 6 will provide the results in quantitative and visual means which will be achieved through the methodology in section 4. Section 7 will discuss and interpret the results whilst providing some limitations. Lastly, section 8 will give the conclusions regarding the extension and potential further research.

2 Background

This section will give the necessary background knowledge to perform and understand the VPint2 extension. This section will go into more details regarding the Sentinel-2 observations, cloud obstructions, cloud detection/masking, the SEN12MS-CR-TS dataset, the VPint2 method and lastly the multi-temporal extension of the VPint2 method.

2.1 Earth Observation and Remote Sensing

Satellite remote sensing has become an important technology for both observing and monitoring Earth's surface. Through such observational systems, we obtain large volumes of imagery which support applications ranging from agriculture to disaster management. A well-known Earth observation is the Sentinel observation mission, operated by the European Space Agency (ESA). Specifically, Sentinel-2 provides multi-spectral high-resolution imagery [6]. These images come from Sentinel-2A, Sentinel-2B and even a more recent Sentinel-2C. The latter of which is not used in the dataset for our work due to the fact that the dataset used in this work was published prior to the launch of Sentinel-2C. Figure 1 shows the descending orbital paths of both Sentinel-2A and Sentinel-2B. Sentinel-2 captures information in thirteen spectral bands spanning the visible, near-infrared, and shortwave infrared regions of the electromagnetic spectrum. The bands provide information about vegetation and soil composition. For these reasons, Sentinel-2 is extremely valuable for optical remote sensing applications.

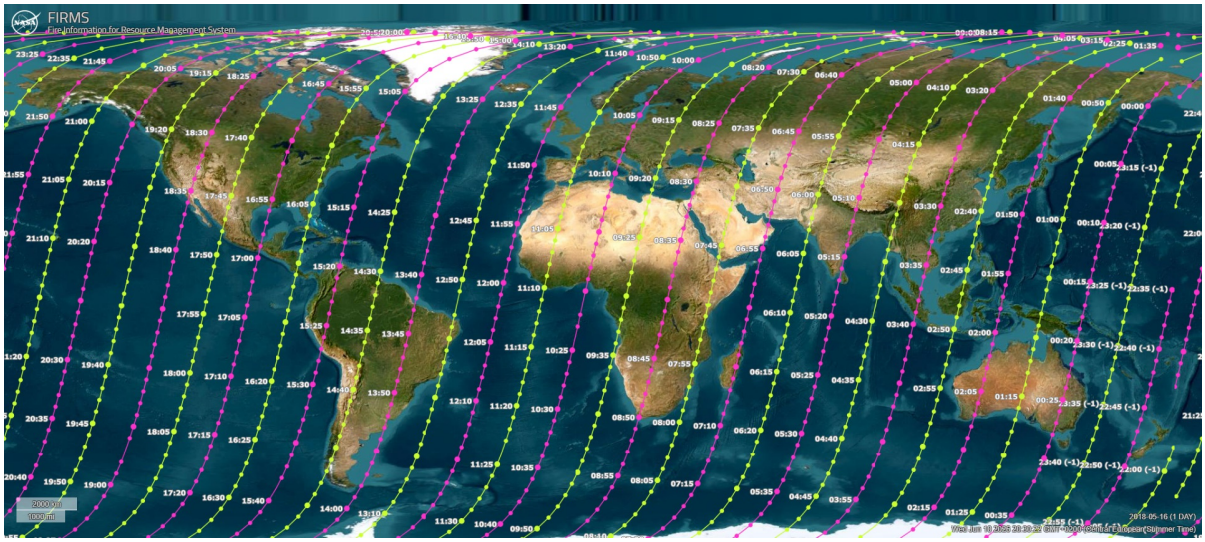


Figure 1: Sentinel-2A's descending orbital path (pink) and Sentinel 2B's descending orbital path (lime green) on 16-05-2018 [7]

Even with all of the advantages of Sentinel-2, some limitations do exist. The observations made by Sentinel-2 can be obstructed by clouds and the shadows of clouds. Cloud obstruction reduces the usability of these images for monitoring and observation-based applications. In some areas, the entire satellite image can be concealed. Thus, cloud obstruction is a major challenge in the field of optical remote sensing.

2.2 Cloud Obstruction

Unlike Synthetic Aperture Radar (SAR), which can penetrate cloud cover due to its microwave signal emission [8], the optical sensors in Sentinel-2 rely on reflected sunlight. Thus, thick clouds can obstruct parts of Earth's surface. The obstruction does not merely end with the clouds themselves, the shadows produced from clouds perturb nearly as much as the clouds obstruct themselves. As mentioned in Section 1, on average 67% of Earth's surface is covered by clouds at any time frame [3]. This limitation is especially problematic for applications that require frequent monitoring of regions on Earth such as for active forest fire monitoring.

The reduction in visibility due to cloud obstruction directly causes a reduction in usability of the applications. Moreover, machine learning models that are trained on the satellite images can start producing unreliable predictions in the presence of cloud coverage. The simplest solution for this matter would be to discard the cloudy observations. The obvious problem with this would be the drastic reduction in sample size of the usable dataset, causing further temporal gaps. Furthermore, the inability to have input data for a given application at inference time and thus being unable to make predictions on missing data, could be catastrophic. Instead, in this thesis, we will focus on ways to make use of satellite images regardless of the cloud coverage.

2.3 Cloud Detection and Cloud Masking

Before the cloud-covered pixels can be reconstructed, they must be identified first. This process of identifying cloudy pixels is referred to as cloud detection or cloud masking [4]. A state-of-the-art

cloud detection model is the SEnSeIv2 [9] model. It must be noted that SEnSeIv2 is treated as a preprocessing step. The cloud detection itself is not the focus of the research. Rather, it provides the cloud masks which are needed for the cloud removal. SEnSeIv2 is a sensor-independent cloud and shadow detection framework which is capable of generating cloud masks across different sensors. Cloud masking is a crucial part of the cloud reconstruction process. The ability to identify which pixels are obstructed and which are cloud-free will form the basis for producing a synthetic reference image alongside running the VPint2 reconstruction.

2.4 Value Propagation Interpolation (VPint2)

Value Propagation Interpolation 2 (VPint2) was created as a training-free cloud removal framework for Sentinel-2 imagery [4]. Contrary to deep learning approaches, VPint2 does not require model training or large annotated datasets. Rather, reconstruction is achieved through interpolation which is guided by the spatial relationships within the selected reference image. The primary assumption of VPint2 is that local spatial relationships remain relatively consistent between both a cloudy target image and a cloud-free reference image. VPint2 exploits this in order to estimate the interpolation weights from the reference image and then propagate the information from observed regions into the cloud-covered areas. Pixel values are iteratively updated based on neighboring observations. Information is propagated from valid pixels towards missing regions until convergence is reached.

The reconstruction that is produced from this preserves local spatial structure whilst simultaneously maintaining computational efficiency. The major advantage of VPint2 is the fact that it maintains its training-free nature. The method itself avoids training costs due to the fact that no model fitting is required. VPint2 was shown to achieve competitive results and performances in comparison to learning-based baselines [4]. VPint2 avoids the large upfront computational cost that deep learning methods require. The primary limitation of the VPint2 method remains the assumption that a cloud-free reference image corresponding to the target scene is available. This assumption does not always hold, especially in regions and seasons with persistent large cloud coverage. This limitation is the primary motivation behind this thesis's research scope.

3 Related Work

This section places the thesis within the broader literature regarding cloud removal. We will discuss spatial methods and their approach(es) to solving the cloud removal problems. Furthermore, the thesis will focus on the discussion regarding multi-modal methods. Relevantly, to the thesis, the section will also mention multi-temporal approaches alongside time-series modeling.

3.1 Spatial Methods

Spatial methods rely on the patterns within the cloud-free regions of an image in order to reconstruct the cloudy regions [4]. Spatial methods often employ convolutional neural networks (CNNs) in order to capture local spatial structure without being explicitly limited to this approach. The main problem is that many such methods suffer from poor scalability. Shen et al. [10] found that interpolation approaches are mainly useful and effective when filling small gaps. In contrast,

downstream cloud removal applications can have clouds that potentially obstruct a relatively large part of the image itself. In combination with the high resolution of the images themselves, interpolation methods can struggle to remain successful in such applications.

3.2 Multi-Modal Methods

Multi-modal methods, on the other hand, primarily exploit the information between different sensors. The most notable of the sensors is the synthetic aperture radar (SAR). SAR signals are largely unaffected by cloud cover and can thus effectively reconstruct the cloud-free image. One of the most important examples of a multi-modal method is the deep learning approach DSen2-CR by [11]. DSen2-CR uses CNNs and leverages SAR data. More recently, diffusion-based approaches have also emerged for SAR-assisted cloud removal. DMDiff [12] uses a dual-branch architecture to extract features from SAR and optical imagery. This is done prior to the fusion of both through a cross-attention mechanism. DMDiff utilizes the iterative denoising process of the diffusion models in order to reconstruct cloud-covered regions. The primary challenge behind such approaches remains the temporal shift. Moreover, SAR data often requires complex preprocessing pipelines, which in turn could produce additional challenges [4].

3.3 Multi-Temporal Methods

For this thesis, the most interesting of the methods are the multi-temporal methods. Multi-temporal methods exploit the temporal information from co-located images acquired at different times to fill data gaps. It is important to not confuse this with time-series modeling methods (the next subsection will go into more detail for time-series modeling methods). A common approach of this nature is called mosaicking [4]. Within this mosaicking approach, users would use cloud masking and then proceed to fetch suitable non-cloudy pixels and mosaic them onto a target image. Such an approach is commonly referred to as temporal replacement. Some relevant practical uses of this type of method include ecological monitoring [13] and mapping [14]. The main limitation of these models is that the accuracy of the reconstructed images is heavily reliant on the availability of suitable cloud-free pixels and the close temporal proximity required to form high quality reconstructions.

3.3.1 Time-Series Modeling Methods

Time-series modeling methods are a specific subset of multi-temporal methods in which the full time-series of normal temporal acquisitions is exploited. In recent years, multiple neural network-based approaches have been proposed for multi-temporal applications, in many cases also incorporating elements of multi-modal methods. UnCRtainTS [15] combines SAR optical data fusion with a multi-temporal approach. This thesis will use multiple approaches and existing methods evaluated on the same dataset, including: DSen2-CR [11], STGAN [16], U-TAE [17] and UnCRtainTS [15]. STGAN is a trainable spatiotemporal generator network designed for cloud removal in optical satellite imagery.

3.4 Summary and Research Gap

To summarize, spatial methods are largely limited when cloud gaps are large. Multi-modal methods on the other hand, often depend on SAR data alongside complex cross sensor preprocessing. Many time-series based methods are deep learning methods which often require large labeled training data and computation. Training-free multi-temporal methods, which uses the data-efficiency of spatial interpolation methods and the temporal richness from time-series modeling methods, remain relatively unexplored. In this thesis, we aim to bridge this gap by extending the VPint2 method to operate over entire satellite image time-series without requiring model training or a guaranteed single fully cloud-free reference image.

4 Methodology

In this work, we will present a detailed and structured methodology which explains how results were obtained. This ranges from the motivation of the extension to the evaluation metrics used to evaluate the performance.

4.1 Motivation Behind The Multi-Temporal Extension

A natural solution to the inherent mono-temporal nature of the VPint2 method was to use the available time-series as a whole. Rather than relying on a single cloud-free observation, information could potentially be aggregated across multiple partially cloudy observations. Within multi-temporal settings, different observations can contain clouds in different locations. This means that some pixels that are obstructed by clouds in one image may be completely cloud-free in another image. By combining information about the pixels from different time steps, it becomes possible to construct a synthetic reference image that approximates the exact cloud-free reference image which VPint2 assumes to be available. The primary logic lies behind the fact that on average it is far more likely to find a cloud-free observation per pixel in the time-series than it is to find a single fully cloud-free reference image in the time-series.

The main challenge in the approach lies in the natural differences among the images across the time-series, primarily due to the potentially large temporal distance between the cloud-free observations per pixel. To combat this, the thesis will incorporate a temporally weighted average in the construction of the synthetic reference image to minimize temporally caused differences within the images. In the general method of this thesis, cloud masks will be used to identify valid observations across the time-series. After which, cloud-free observations are averaged in order to construct a synthetic reference image. The construction of the synthetic reference image will be weighted temporally to favor observations temporally closer to the target image. Subsequently, the resulting synthetic reference image will be used within the original VPint2 method as a cloud-free reference image. Through this, the VPint2 method will be extended from a mono-temporal to a multi-temporal reconstruction framework.

4.2 Cloud Detection

As mentioned in Section 2, the cloud detection is treated as a prerequisite step which is performed once per ROI (Region Of Interest) and patch. The SEnSeIv2 model [9] is chosen as the model used

for per-pixel cloud classification, though this step is a preprocessing step which can be performed by any cloud detection model/algorithm. When the model is called, it outputs a discrete class label map, instead of probability scores. The output is a mask whose values are an element of the set {0, 1, 2, 3}. These four values represent four classes in numerical increasing order:

- 0: no clouds and no shadow detected.
- 1: thin clouds present. Surface under the clouds may be visible but biased.
- 2: thick clouds present. No surface underneath the clouds is viewable.
- 3: cloud shadow present.

Let i be the row coordinates and j be the column coordinates of a pixel in the image. We take the $\text{logit}_k(i,j)$ as the model's raw score for a class. In other words, $\text{logit}_k(i,j)$ determines how strongly the model believes pixel (i,j) belongs to class $k \in \{0,1,2,3\}$. If the raw output has three dimensions, we perform an argmax operation in order to recover the single-channel class map:

$$\text{raw mask}(i,j) = \arg\max_k [\text{logit}_k(i,j)] \tag{1}$$

Through the argmax, we obtain the class which the model assigns the highest confidence to for each pixel. Subsequently, from this four-class raw mask, two binary masks are produced. The primary binary mask ($m_t(i,j)$) marks any pixel that is not clear as cloud-affected for a timestep, thus we include all three non-clear classes:

$$m_t(i,j) = \begin{cases} 1, & \text{raw mask}(i,j) > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

It is important to note that raw mask > 0 is true for any non-clear class and produces a binary mask. A second cloud-only mask is also produced which isolates thin and thick cloud pixels (excluding shadows):

$$\text{cloud only}(i,j) = \begin{cases} 1, & \text{raw mask}(i,j) \in \{1,2\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

The binary mask is used for all reconstruction and evaluation steps for our extended model in this work. Since thin clouds, thick clouds and shadows all prevent reliable surface observation, a binary mask was preferred over a cloud-only mask. The cloud-only mask was kept as a diagnostic in order to confirm that the mask works as expected. Moreover, it provides useful visualization for further understanding the structure of the cloud obstructions within the satellite images. The cloud coverage was computed for a specific time step as the mean of the binary mask. This gives the fraction of pixels that are either cloudy or covered by shadows in some way. Figure 2 shows an example of the cloud masking/construction.

4.3 Synthetic Reference Image Construction

Arriving at the central part of the methodology and the core part of the thesis. In the original VPint2 method, a single fully cloud-free reference image was assumed to be available [4]. It will

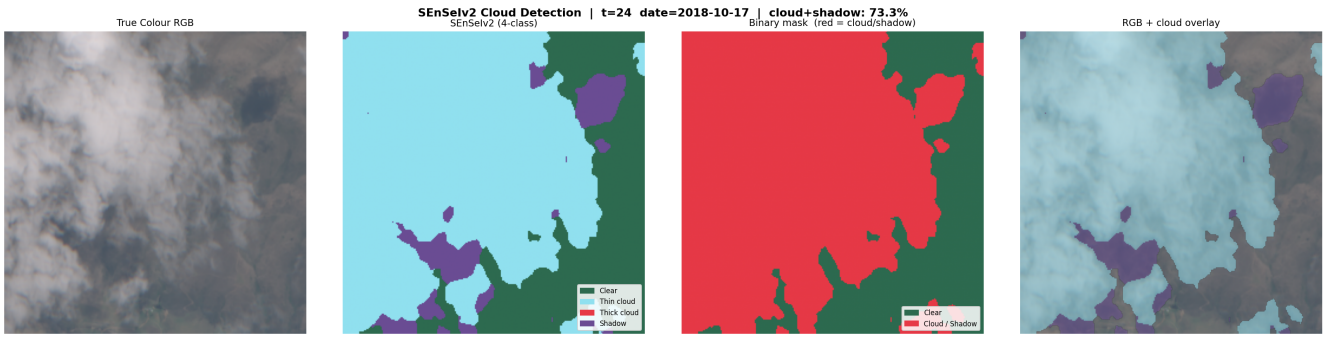


Figure 2: Cloud masking for a specific time step. The cloud detection shows the pixels that are clear, have thin clouds, thick clouds or some form of shadow in the second panel. The figure also shows the binary masks where any non-clear pixel (i.e. binary mask$(i, j) > 0$) is marked as red in the third panel. Lastly, the 4th panel includes an overlay with the raw mask and the target image.

be important to implement an extension which is independent of the existence of a single fully cloud-free reference image, given that the SEN12MS-CR-TS dataset does not guarantee such a single fully cloud-free reference image [5]. The proposed solution in this thesis revolves around constructing a per-target synthetic reference image from the remaining $T - 1$ observations. The solution makes use of temporal proximity weights.

4.3.1 Synthetic Weight Assignment

All the time steps other than the target t are labeled as candidate observations. For each candidate time step $t' \neq t$, a temporal proximity weight is computed:

$$w(t') = \frac{1}{|t' - t| + \epsilon} \tag{4}$$

Where $\epsilon = 10^{-6}$ is a constant to avoid any division by 0 in case an error in the dataset exists where two time steps have the same indices. Intuitively, Equation 4 demonstrates that the weighting decreases as the temporal gap between the target image and observation increases. In other words, adjacent time steps receive the highest weighting in the construction of the time-step-specific synthetic reference image. We also compute the logical complement of the binary mask:

$$v(t', i, j) = 1 - m_{t'}(i, j) \tag{5}$$

Where $m_{t'}(i, j)$ is the binary cloud mask for all candidate time steps. We then compute the effective per-pixel weight for candidate t':

$$w_{\text{eff}}(t', i, j) = w(t') * v(t', i, j) \tag{6}$$

Through this, we assign a zero weight to cloudy pixels at any given candidate time step and location. Clear pixels, on the other hand, will be weighted by temporal proximity.

4.3.2 Weighted Reference Aggregation

The synthetic reference image is computed as the weighted mean of the candidate observations. Let $x_{t'}(c, i, j) \in [0, 1]$ denote the normalized reflectance of spectral band c at pixel (i, j) in time step t':

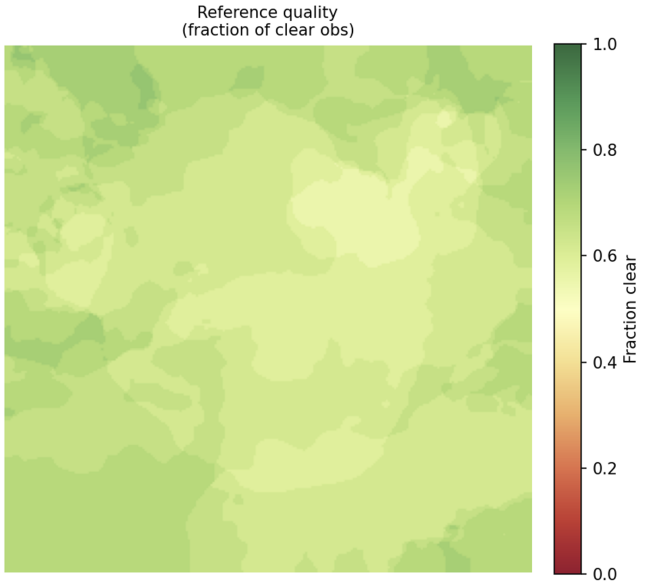
$$r_t(c, i, j) = \frac{\sum_{t' \neq t} w_{\text{eff}}(t', i, j)\, x_{t'}(c, i, j)}{\sum_{t' \neq t} w_{\text{eff}}(t', i, j)} \tag{7}$$

Where weighting is determined by w_{eff}. The denominator is the total effective weight accumulated across all the candidate time steps for pixel (i, j). Figure 3 shows an iteration of the reconstructions that our model produces whilst using the constructed synthetic reference image. Figure 4 shows the associated VPint2 reconstruction.

(a) Target image with the calculated cloud coverage percentage (19.9% in this case).

(b) Binary cloud mask where each non-clear pixel is marked as red and each clear pixel is marked as green.

(c) Synthetic referenced image constructed based on the given timestep of the target image.



(d) Reference quality map where greener patches indicate a high fraction of clear observations for those pixels and red indicates lower fraction of clear observations

Figure 3: Example of a target image (a) with the corresponding binary cloud mask (b), synthetic reference image (c), reference quality image (d).

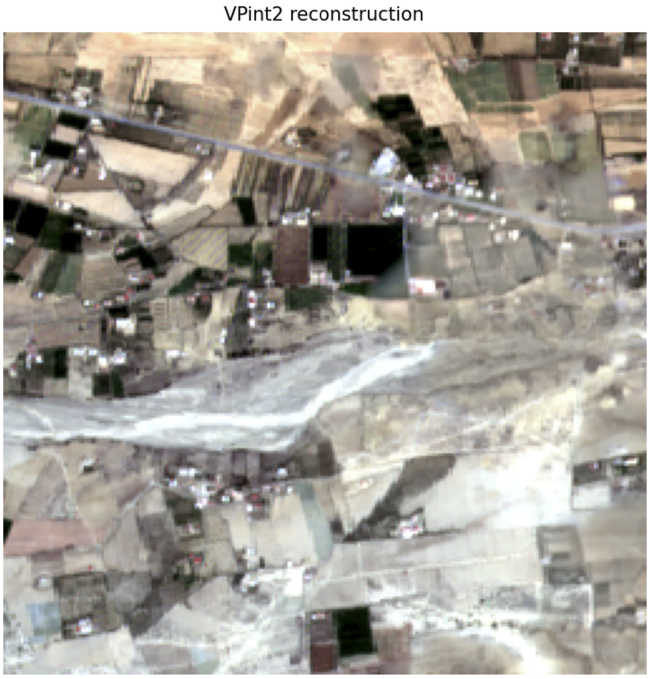


Figure 4: Vpint2 reconstruction based on the timestep in Figure 3

When a pixel has no cloud-free observation across any of the $T-1$ candidate time steps in the time-series, a median value is taken. If $\sum_{t'} v(t', i, j) = 0$, the approach will take the temporal median value of all candidate pixel observations, regardless of the cloud status on the pixel:

$$r_t(c, i, j) = \text{median}_{t' \neq t}(x_{t'}(c, i, j)) \tag{8}$$

Within this fallback, the temporal median is preferred over the mean since it will be more robust to spectral outliers. Alongside the reference image, a quality map is also produced for the reference image itself. The quality value at each individual pixel is determined by the fraction of candidate time steps at which the pixel was observed to be cloud-free:

$$Q_t(i, j) = \frac{1}{T-1} \sum_{t' \neq t} v(t', i, j) \tag{9}$$

Figure 3d shows the reference quality image. In that particular case, we see a high fraction of clear observations for most pixels in the image. This implies that the model had many observations to sample from in order to form a temporally weighted average.

4.4 VPint2 Reconstruction

Now that we have the synthetic reference image r_t and the cloud mask m_t, we can apply the original VPint2 method [4] in order to reconstruct the cloud-affected pixels of the target image x_t.

4.4.1 Target Image Preparation

For each spectral band c, we take the target band array and initialize it as a copy of the observed target image band $x_t(c)$.

$$\tilde{x}_t(c,i,j) = \begin{cases} \text{undefined}, & m_t(i,j) = 1 \\ x_t(c,i,j), & m_t(i,j) = 0 \end{cases} \tag{10}$$

The corresponding reference image is already normalized to the [0,1] reflectance range during preprocessing. The data from Sentinel-2 is originally provided as values in the range of [0,10000]. These are scaled by dividing by 10000 and clipped to [0,1] as a part of the data loading step. Prior to being passed to VPint2, the reference image is clipped again to [0,1] as a safeguard. This safeguard ensures that the numbers remain stable and prevents the production of out-of-range reflectance values.

4.4.2 WP-MRP Weight Propagation

The WP-MRP algorithm operates by iteratively updating each missing pixel value based on a weighted combination of its spatial neighbors. The weights are computed analytically from the reference image values, without needing any learned model. Unlike the original VPint algorithm, no reliance of machine learning approximation is needed. The update rule of the algorithm in question for each missing pixel (i,j) is:

$$\hat{x}(i,j) = \Sigma_{n \in N(i,j)}[\alpha_n * \hat{x}(n)] \tag{11}$$

$N(i,j)$ is the set of all spatial neighbors of the pixel (i,j). α_n are the predicted propagation weights. These weights come from the spatial structure encoded in the reference image. Pairs of the adjacent free pixel values are used as the input feature.

The VPint2 algorithm also handles two notable edge cases. If the cloud coverage of a target image is 0%, the reconstruction step will correctly be skipped and the original target image will be returned. On the contrary, if a time step has a cloud coverage of 100%, then the reconstruction step will simply return the synthetic reference image in its entirety.

5 Experiments

The fundamental aspects of the experimental setup that we used are essential to answering the research question. In order to produce a multi-temporal cloud removal to the time-series setting, we first introduced the SEN12MS-CR-TS dataset [5]. Moreover, evaluating the extension that the research method mentions requires existing published models. These models are also introduced in this section.

5.1 The SEN12MS-CR-TS dataset

In order to research cloud removal, multiple benchmark datasets have been developed. One such dataset is SEN12MS-CR-TS [5]. SEN12MS-CR-TS provides multi-modal and multi-temporal observations, which will be useful for cloud removal research and applications. SEN12MS-CR-TS provides time-series of both Sentinel-1 SAR imagery and Sentinel-2 optical imagery collected over various geographical regions. For the scope of this thesis, the Sentinel-1 SAR data will be ignored.

The primary reason behind this is that VPint2 was tailored towards the Sentinel-2 data. The major advantage of this dataset with respect to the scope of the thesis is the realistic cloud distributions it provides. Moreover, the SEN12MS-CR-TS dataset is very commonly used to compare methods. Likewise, this dataset is also commonly used as a benchmark against other relevant methods. This allows for robust and valid comparisons to be made which are present in Section 6.

This dataset also provides sufficient temporal depth which enables the exploration of multi-temporal reconstruction strategies and approaches. Additionally, the SEN12MS-CR-TS dataset has been used by many other models within cloud removal literature. This allows for accurate comparisons between the extension of VPint2 and other published models such as DSen2-CR [11], STGAN [16], U-TAE [17] and UnCRtainTS [15].

5.2 Time-Series Loading and Normalization

Upon downloading the SEN12MS-CR-TS dataset, the raw satellite data are stored on disk as GeoTIFF files. Each patch consists of 13 spectral bands at 256×256 pixels at 10m resolution. All of the bands are read at the same time into a NumPy array of shape $(C, H, W) = (13, 256, 256)$. Pixel values are stored as 16-bit unsigned integers with a fixed factor of 10000. All pixel values are stored as Level-1C top of atmosphere reflectance integers which are scaled by a factor of 10000 and are subsequently normalized to a [0,1] range prior to the processing:

$$\text{x norm} = \text{clip}(\frac{\text{x raw}}{10000}, 0, 1) \tag{12}$$

The normalization in Equation 12 ensures that all spectral values are expressed as reflectances within the valid range. This preprocessing step is applied because SEnSeIv2 expects reflectance inputs in the [0,1] range. Such a normalization also aids in the processing and metric evaluation by ensuring that all spectral bands are consistent in their numerical range. The clipping is also essential to remove physically invalid reflectance values which may be present due to some noise. Although VPint2 is robust to input ranges, the normalization is still kept and preserved throughout the entire pipeline to provide consistency. The entire time-series is assembled through the stacking of the multispectral images along a new leading axis $(T, 13, 256, 256)$ with $T \leq 30$.

5.3 Quantitative Evaluation

In this experiment, we will perform quantitative evaluation in order to compare the reconstructed image against ground truth observations. Five evaluation metrics are used: RMSE, MAE, PSNR, SSIM and SAM for both the complete regions and only the cloudy regions. In this thesis, we will compare our results to the results of other state-of-the-art baselines.

- DSen2-CR is a deep learning model capable of removing clouds from Sentinel-2 images [11]. In this method SAR-optical data fusion is used in order to exploit the synergistic properties of the images. This then guides the image reconstruction alongside a cloud-adaptive loss to maximize the preservation of the original information.

- STGAN is a trainable spatiotemporal generator network, which serves to remove clouds [16]. STGAN outperforms many standard models and has a high PSNR and SSIM scores across a variety of atmospheric conditions.

- U-TAE (U-net with Temporal Attention Encoder), is a novel spatio-temporal encoder which combines multi-scale spatial convolutions alongside a temporal self-attention mechanism [17].
- UnCRtainTS is a method for multi-temporal cloud removal. This method combines a novel attention-based architecture alongside a formulation for multivariate uncertainy prediction [15].

5.3.1 Ground Truth Selection

In order to make any meaningful comparison, a suitable ground truth must be chosen. This thesis follows a specific and strict methodology for determining the most appropriate ground truth to use [15]:

1. If any time step after the target has a cloud coverage below the cloud-free threshold (default at 10%), then the nearest cloud-free timestep under the threshold is chosen. In other words, we produce a set of all time steps after the target whose cloud coverage lies below the threshold. Of that set, the nearest time step is chosen (with respect to the target time step). This allows for temporally close ground truths whilst still ensuring that cloud coverage is kept to a minimum.
2. If no time step after the target meets the 10% threshold, then the least cloudy time step among all the later time steps is chosen to be ground truth. This fallback serves as a way to attempt to minimize the cloud coverage of the ground truth regardless of the time-series.
3. If the target time step is the final time step in the entire time-series, then the search is reversed. In other words, steps 1 and 2 are repeated with the only difference being that the focus lies in the preceding time steps rather than the proceeding time steps.

The 10% cloud-free threshold was chosen due to the good tradeoff it offers. A threshold set too high will almost guarantee a temporally close ground truth. However, it might be susceptible to ground truths with potentially high cloud coverage. Similarly, thresholds set too low will focus on potential ground truths with little cloud coverage. However, the risk is now that the temporal distance can become large. For those reasons, 10% was seen as a good tradeoff for the ground truth. This offered a balance between attempting to obtain a temporally close ground truth and attempting to have little cloud coverage of the ground truth.

5.3.2 Evaluation Metrics

Let $\hat{x} \in \mathbb{R}^{C \times H \times W}$ denote the reconstruction and $g \in \mathbb{R}^{C \times H \times W}$ be the ground truth image. Where $C = 13$ is the number of spectral bands, $H = 256$ and $W = 256$ are the spatial dimensions. We have Ω as either the full pixel set or the set of spatial pixel positions determined as cloudy by the binary-mask (in Equation 2).

Starting off with Root Mean Square Error (RMSE). RMSE is computed as the square root of the mean squared error per pixel, which is averaged initially across the $C = 13$ spectral bands:

$$\text{RMSE} = \sqrt{\frac{1}{|\Omega|}\Sigma_{(i,j)\in\Omega}\frac{1}{C}\Sigma_{c=1}^{C}(\hat{x}_{c,i,j} - g_{c,i,j})^2} \tag{13}$$

Secondly, we compute Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{|\Omega|}\Sigma_{(i,j)\in\Omega}\frac{1}{C}\Sigma_{c=1}^{C}|\hat{x}_{c,i,j} - g_{c,i,j}| \tag{14}$$

MAE is less sensitive to large outlier errors in comparison to RMSE. Thus, it can be useful as a complementary measure.

Thirdly, this thesis will use PSNR (Peak Signal-to-Noise Ratio):

$$\text{PSNR} = 10\log_{10}(\frac{MAX^2}{MSE}), MAX = 1 \tag{15}$$

In Equation 15, $MAX = 1$ is the maximum possible reflectance value. MSE is the mean squared error averaged across bands and pixels in Ω.

Fourthly, we make use of structural similarity index measure (SSIM). SSIM deploys a modified measure of spatial correlation between the pixels of a reference alongside test images in order to quantify the degradation of the structure of an image [18]. In the equation below, we have $\mu_{\hat{T}}, \mu_T$ as local means, $\sigma^2_{\hat{T}}, \sigma^2_T$ as local variance. We have that $\sigma_{T\hat{T}}$ is the local cross-covariance and lastly $c1$ and $c2$ are stability constants[19].

$$\text{SSIM}(\hat{T}, T) = \frac{(2\mu_{\hat{T}}\mu_T + c_1)(2\sigma_{\hat{T}T} + c_2)}{(\mu^2_{\hat{T}} + \mu^2_T + c_1)(\sigma^2_{\hat{T}} + \sigma^2_T + c_2)} \tag{16}$$

Lastly, the thesis incorporates SAM [20]. SAM takes the arccosine of the dot product of the spectra. SAM determines the similarity of a test masked spectrum to a reference spectrum.

$$\text{SAM}(i,j) = \arccos\left(\frac{\hat{\mathbf{x}}^{\top}_{i,j}\mathbf{g_{i,j}}}{\|\hat{\mathbf{x}}_{i,j}\|_2 \cdot \|\mathbf{g}_{i,j}\|_2}\right) \tag{17}$$

$$\text{SAM} = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}\text{SAM}(i,j) \tag{18}$$

Beyond using the common evaluation metrics, we will also look at reference quality and reconstruction errors. The thesis computes a reference quality chart. This chart portrays the fraction of candidate time steps where pixel (i, j) was cloud-free. The thesis will also compute an absolute MAE per pixel (only averaged out over the different bands per pixel) error map as shown in Figure 5:

$$e(i,j) = \frac{1}{C}\Sigma_{c=1}^{C}|\hat{x}_{c,i,j} - g_{c,i,j}| \tag{19}$$

Subsequently, the error map $e(i, j)$ is correlated against the reference quality map Q_t using both Pearson's r and Spearman's ρ, as shown in Figure 6.

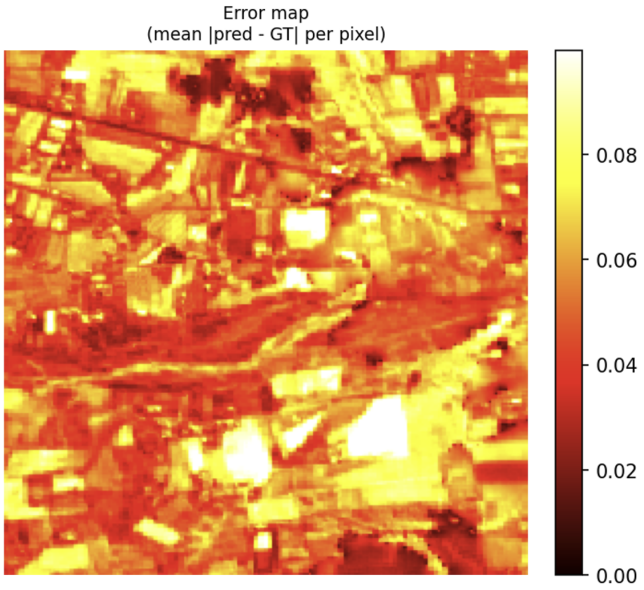


Figure 5: Error map comparing the MAE per pixel (taken as the mean MAE per band per pixel) for different regions of the reconstruction. Darker regions imply lower MAE whereas brighter regions imply a higher MAE.

6 Results

In this section, we present the quantitative results obtained through following the methodology alongside conducting experiments using the multi-temporal time-series extension of the VPint2 method. First, the performance of our extended model was evaluated across the different geographical regions included within the Sentinel-2 SEN12MS-CR-TS dataset. Subsequently, the performance of the entire extended model will be compared against existing state-of-the-art methods. Moreover, our extended VPint2 model must be evaluated prior to being used to answer the research question.

6.1 Comparison with Existing Baselines

In Table 1, we observe the performance of the VPint2 model that we extended against other existing methods. Overall, the extended model performed very competitively and even outperformed the other methods in some metrics. The results do indicate that the original VPint2 method outperformed the extension that we introduced on all metrics on a substantial level. Though such interpretations based on the observations must be made carefully and in perspective. The original VPint2 method was evaluated on a suitable subset of Sentinel-2. The extended VPint2 was evaluated on the vast majority of the Sentinel-2 portion of the SEN12MS-CR-TS dataset. This entire portion included regions that the extended model found difficult, such as Europe and East Asia, as shown in Tables 2 and 3. The performance of the extended VPint2 model relative to the other existing methods varied across the different metrics. Our model performed poorest in RMSE yet outperformed many other existing models (besides the original VPint2) in PSNR and SAM. Similarly, the extended model outperformed most existing methods in the SSIM metric.

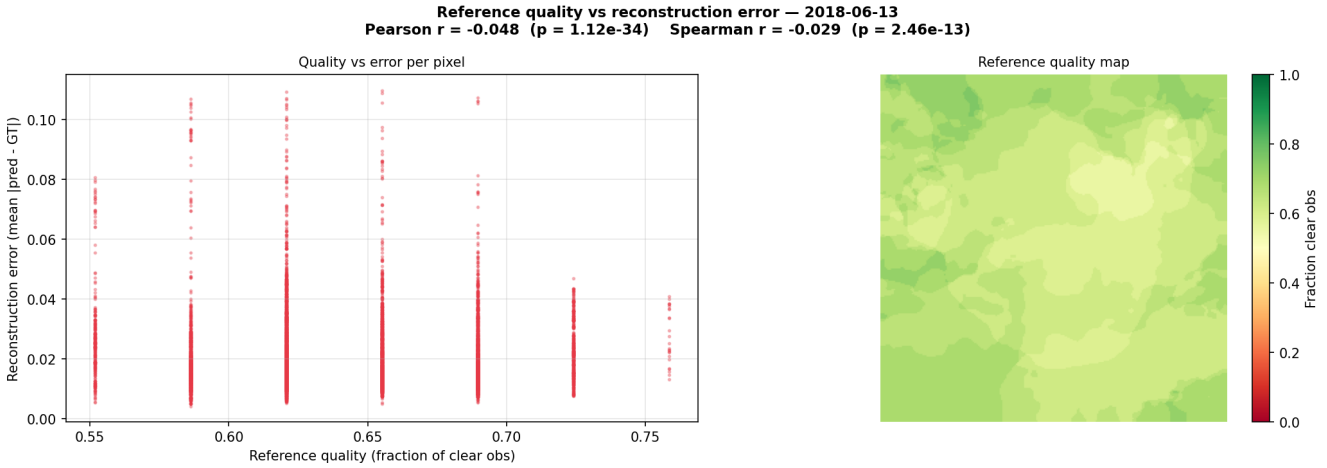


Figure 6: Scatter plot showing a downwards trend in error as the reference image has larger fractions of clear observations. To the right, we obtain a reference quality image showing the fraction of clear observations per pixel. Observations are supported by the Pearson's r, Spearman's ρ and p-values.

Table 1: Comparing the performance results from the multi-temporal methods on the full SEN12MS-CR-TS dataset [15] to the results of the extension of the VPint2 method. The extension that was presented in this work was evaluated on approximately 90% of the Sentinel-2 dataset within the SEN12MS-CR-TS dataset. The original VPint2 method was evaluated on a suitable subset of the Sentinel-2 dataset [4]. The presented extension of VPint2 is computed as a weighted average across all the continents of the Sentinel-2 dataset (weighted by representation based on the total number of evaluated patches per region) for both the cloudy/shadow only region and the entire reconstruction. This approach is equivalent to taken all the instances of the experiments and piling them together to compute an average. The table includes the performances of the non-VPint2 baselines [15] and the original VPint2 results [4]

Methods	RMSE (↓)	PSNR (↑)	SSIM (↑)	SAM (↓)
DSen2-CR	0.079	26.04	0.810	12.147
STGAN	0.060	25.42	0.818	12.548
CR-TS Net	0.057	26.68	0.836	10.657
U-TAE	0.051	27.05	0.849	11.649
UnCRtainTS	0.051	27.84	0.866	10.160
Original VPint2 (uses suitable subset of data)	**0.042**	**30.38**	**0.928**	**6.541**
Extension VPint2 Entire Reconstruction (ours)	0.085	27.80	0.838	9.050
Extension VPint2 Cloudy Regions Only (ours)	0.086	27.61	0.838	9.161

6.2 Regional Performances of the Extended Model

Table 2 shows varying results regarding the performance of our extended model relative to the different regions.

Table 2: Evaluating all the different regions/continents within the Sentinel-2 SEN12MS-CR-TS dataset [5]. The evaluation metrics are RMSE, MAE, PSNR, SSIM and SAM. Evaluations are performed over the entire VPint2 reconstruction image regardless of cloudy/shadow regions in the target image. Our extended VPint2 model is run on the majority of Sentinel-2 but not the entire dataset.

Regions of Experimentation	RMSE (↓)	MAE (↓)	PSNR (↑)	SSIM (↑)	SAM (↓)	Total Patches
Africa	0.0410	0.0338	32.2832	0.9265	5.5867	1960
Americas	0.0904	0.0783	27.7236	0.8417	7.5646	2203
East Asia	0.1091	0.0945	26.2637	0.8185	8.8205	2277
Europe	0.1178	0.1010	23.5629	0.7607	13.3400	2509
West Asia	0.0548	0.0453	30.6615	0.8674	8.9612	1920

The extended model on the region Africa achieved by far the strongest performance across all the five evaluation metrics. On the contrary, the region Europe had the worst score across all of the metrics. In terms of the RMSE score, VPint2 performed nearly three times worse on Europe in comparison to Africa. One metric where the region Europe performed particularly poorly relative to the other regions is the SAM metric. The regions Americas and East Asia fall in the middle of the table with West Asia having a similarly strong performance to Africa.

Table 3 shows the performance of the extended model for the different regions over only the pixels that are cloudy or covered by cloud shadows in the target image. Overall, the results across the different regions for the different metrics are nearly identical to the results in Table 2. Consequently, the patterns and variations with the regions remain the same. This indicates that the reconstruction quality of VPint2 remains relatively consistent regardless of the region that the reconstruction is being evaluated on. A noticeable difference between the results both in Table 2

Table 3: Evaluating the different regions/continents within the majority of the Sentinel-2 SEN12MS-CR-TS dataset [5]. The evaluation metrics are RMSE, MAE, PSNR, SSIM and SAM. Evaluations are performed only over the regions of the VPint2 reconstruction image which are covered with clouds/shadows in the target image.

Regions of Experimentation	RMSE (↓)	MAE (↓)	PSNR (↑)	SSIM (↑)	SAM (↓)	Total Patches
Africa	0.0420	0.0347	32.0276	0.9265	5.7516	1960
Americas	0.0912	0.0792	27.4439	0.8417	7.6668	2203
East Asia	0.1093	0.0949	26.1713	0.8185	8.8757	2277
Europe	0.1181	0.1015	23.4681	0.7607	13.4469	2509
West Asia	0.0557	0.0462	30.4067	0.8674	9.0921	1920

and 3 is the fact that the model slightly under performs in Table 3. It must be noted that the clear pixels in the full image reconstruction largely carry non-zero errors. This is shown by the error map in Figure 5 with the appropriate cloud masking in Figure 3b. The primary reason behind this lies in the fact that the ground truth differs from the target image and thus slight temporal difference can arise even in the clear pixels. Another notable observation is the fact that the results for the SSIM metric remained identical across both Table 2 and 3. The reason behind this is that the current experiment was set up to use the SSIM metric over the entire reconstruction. This element and the usage of SSIM will be evaluated in Section 7.

7 Discussions and Limitations

The initial notable discrepancy between the results was the regional variations which were present in both the entire reconstruction and the cloudy/shadow only evaluation results. A likely explanation behind these regional variations in performance is due to landscape changes. As shown in Tables 2 and 3, Africa and West Asia were the best performing regions. Both Africa and West Asia tend to have large homogeneous regions such as savannas and deserts. Such regions likely present simpler spectral and temporal patterns with fewer variations as time goes on. Regions that performed poorly, like Europe, on the other hand, have many diverse landscapes. Such landscapes range from dense urban areas to isolated forests and mountains. These landscapes tend to have more dynamic seasonal changes, this in turn could make temporal interpolation harder. The relatively high SAM results observed for Europe suggest spectral shape is recovered poorly in such regions. This may imply fast changes in land cover between the time steps.

Interesting interpretations can also be drawn from the differences, or lack thereof, in the evaluation results for cloudy/shadow regions and those for the entire reconstruction. The fact that the results from Tables 2 and 3 are nearly identical implies that VPint2 does not struggle much with the pixels that matter most for this relevant research. Regarding the comparisons between our extended VPint2 model and the other existing models, our model performed well and competitively. As mentioned prior in Section 6, the original VPint2 approach did outperform our model in all metrics. Though it should still be noted that original VPint2 was performed on a suitable subset of Sentinel-2 whereas our model was performed on nearly the entire Sentinel-2 dataset. Moreover, the weighted average (weighted by representation based on the total number of patches per region) of our extended model for both the cloudy and entire region performed well across the metrics. Most notably, our extended model outperformed all non-VPint2 approaches in PSNR and SAM metrics. Regarding RMSE and PSNR, both of which are computed on the same per-image MSE. The values inside the tables are averaged separately across the reconstructions. That means that the mean PSNR is not equivalent to converting the mean RMSE back into PSNR. This suggests that the extension performs reasonably well in comparison to other approaches across diverse conditions and landscapes. The low SAM score implies that spectral signatures are preserved. Good PSNR and SSIM scores would mean that visual and structural quality is competitive, which Table 1 shows. Yet the slightly worse RMSE which was also observed implies some reflectance errors. This implies that the extension of VPint2 might be less useful for applications requiring highly accurate pixel level reflectance values.

7.1 Limitations

Though the results from the extended model were positive, some notable limitations still exist that are worth addressing. One such limitation is the fact that the SSIM score is identical in both the entire reconstruction evaluation and the evaluation on only the cloudy/shadow regions. The main reason behind this was an error in the experimental setup which caused the SSIM value to be computed over the entire reconstruction independent of the cloudy/shadow regions. This is primarily an implementation limitation which would require a more careful approach to fix. Another limitation was the inability of our method to ensure a truly cloud-free ground truth image. Currently, the threshold for an observation within the time-series to be considered cloud-free enough is at 10%. One could argue that such a threshold is quite high and would allow for somewhat cloudy ground truth images. Cloudy ground truth images would partially reduce the validity of the evaluation results due to potential presence of clouds in the ground truth image independent of the VPint2's reconstruction quality. Though a higher threshold would likely ensure a close temporal proximity for the ground truth to the target image, this may not always be needed. As mentioned earlier in this section, some regions such as Africa and West Asia contain relatively homogeneous landscapes already. It is entirely possible for such regions to experience a relatively low amount of visual variation based on temporal differences. Thus, for such regions, it may be more valuable to limit the cloud presence in the ground truth image rather than minimizing the temporal distance of the ground truth to the target image.

Regarding cloud presence, another limitation exists in the way the method handles cases where no cloud-free observations exist for a certain pixel in a time-series during the synthetic reference image construction. Currently, the method selects the median observation regardless of cloud status when constructing the synthetic reference image. A more natural approach could be to produce a synthetic reference pixel with all having an equal VPint weight of one. This would ensure that the cloudy pixel will at the very least be constructed as a type of mean instead of picking a median. Though this may prove to be difficult to achieve with just a reference image. This makes the median approach likely the safest approach in cases where no cloud-free pixel observations exist.

Another limitation lies in the fact that the extended VPint2 model was not run on the entire Sentinel-2 dataset. Currently, approximately 90% was processed with our model, with all the continents (besides Europe) being processed entirely. Europe was the only region where a majority, but not all, of the data was processed. Though running the experiments on the entire Sentinel-2 dataset was planned, technical difficulties proved to be an obstacle. Combined with the fact that Europe is twice as large (in terms of disk space) as any of the other continents. Nevertheless, a vast majority of the Sentinel-2 dataset was processed and results have stabilized. It is unlikely that the evaluation results for our extended VPint2 model would have changed much even with the entire Sentinel-2 dataset.

8 Conclusions and Further Research

This work presented an extension of the original VPint2 approach to the multi-temporal satellite image time-series setting. Our work accomplished this by producing a synthetic reference image which removed the reliance on a single fully cloud-free observation in a time-series. The extension was applied to the majority of the Sentinel-2 dataset across different continents in the world. Through this, we broadened the evaluation relative to the original VPint2 evaluation which worked

only with a suitable subset.

The results in Section 6 demonstrated that our extension performs competitively against existing multi-temporal cloud removal methods. Metrics such as PSNR and SAM were ones where our model performed exceptionally well. This shows the robustness of VPint2 in performing across diverse landscapes. The near identical evaluation results in Tables 2 and 3 implied that VPint2 does not disproportionately struggle at the most important pixels for cloud removal.

On the other hand, in Section 7, several limitations were identified. These limitations can be used to drive further research into this issue. Such research includes reducing the cloud-free threshold for ground truth selection to prioritize cloud-freeness rather than temporal proximity. This especially holds true for relatively homogeneous regions and landscapes. A corrected masked SSIM can also be implemented in order to measure the SSIM for the regions of the VPint2 reconstruction which are cloudy or covered by shadows in the target image. Moreover, a deeper investigation may also be warranted into the convergence behavior of VPint2 across heterogeneous landscapes such as Europe. Convergence diagnostics can be used in the future to test whether optimization behavior contributes to regional differences like in Europe.

References

[1] NASA Land, Atmosphere Near real-time Capability for EOS (LANCE) FIRMS. Fire information for resource management system (firms). https://firms.modaps.eosdis.nasa.gov/, n.d.

[2] NASA Earthdata. Agriculture production. `https://www.earthdata.nasa.gov/topics/human-dimensions/agriculture-production`, n.d.

[3] Michael D. King, Steven Platnick, W. Paul Menzel, Steven A. Ackerman, and Paul A. Hubanks. Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. *IEEE Transactions on Geoscience and Remote Sensing*, 51(7):3826–3852, 2013.

[4] Laurens Arp, Holger Hoos, Peter van Bodegom, Alistair Francis, James Wheeler, Dean van Laar, and Mitra Baratchi. Training-free thick cloud removal for sentinel-2 imagery using value propagation interpolation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216:168–184, 2024.

[5] Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. SEN12MS-CR-TS: A Remote Sensing Data Set for Multi-modal Multi-temporal Cloud Removal. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[6] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote Sensing of Environment*, 120:25–36, 2012. The Sentinel Missions - New Opportunities for Science.

[7] NASA FIRMS. Fire information for resource management system (firms) global fire map. https://firms.modaps.eosdis.nasa.gov/map/, n.d.

[8] Africa Ixmucane Flores-Anderson, Kelsey E Herndon, Rajesh Bahadur Thapa, and Emil Cherrington. The synthetic aperture radar (sar) handbook: Comprehensive methodologies for forest monitoring and biomass estimation. *NASA: Washington, DC, USA*, 2019.

[9] Alistair Francis. Sensor independent cloud and shadow masking with partial labels and multimodal inputs. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–18, 2024.

[10] Huanfeng Shen, Xinghua Li, Qing Cheng, Chao Zeng, Gang Yang, Huifang Li, and Liangpei Zhang. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geoscience and Remote Sensing Magazine*, 3(3):61–85, 2015.

[11] Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

[12] Wenjuan Zhang, Junlin Mei, and Yuxi Wang. Dmdiff: A dual-branch multimodal conditional guided diffusion model for cloud removal through sar-optical data fusion. *Remote Sensing*, 17(6), 2025.

[13] Fengli Zou, Qingwu Hu, Yichuan Liu, Haidong Li, Xujie Zhang, and Yuqi Liu. Spatiotemporal changes and driving analysis of ecological environmental quality along the qinghai–tibet railway using google earth engine—a case study covering xining to jianghe stations. *Remote Sensing*, 16(6):951, 2024.

[14] John Brandt, Jessica Ertel, Justine Spore, and Fred Stolle. Wall-to-wall mapping of tree extent in the tropics with sentinel-1 and sentinel-2. *Remote Sensing of Environment*, 292:113574, 2023.

[15] Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2086–2096, 2023.

[16] Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1796–1805, 2020.

[17] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[18] David M Rouse and Sheila S Hemami. Understanding and simplifying the structural similarity metric. In *2008 15th IEEE international conference on image processing*, pages 1188–1191. IEEE, 2008.

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[20] F.A. Kruse, A.B. Lefkoff, J.W. Boardman, K.B. Heidebrecht, A.T. Shapiro, P.J. Barloon, and A.F.H. Goetz. The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, 44(2):145–163, 1993. Airbone Imaging Spectrometry.

Link to the GitHub Repository of the code used in this paper